

## DOCUMENT RESUME

ED 264 754

HE 018 927

AUTHOR Saavedra, Pedro; And Others  
TITLE Quality Control Analysis of Selected Aspects of Programs Administered by the Bureau of Student Financial Assistance. Task 1 and Quality Control Sample; Error-Prone Modeling Analysis Plan.  
INSTITUTION Applied Management Sciences, Inc., Silver Spring, Md.  
SPONS AGENCY Bureau of Student Financial Assistance (DHEW/OE), Washington, D.C.  
PUB DATE 31 Mar 80  
CONTRACT 300-79-0742  
NOTE 39p.; For related documents, see HE 018 926-928. Task 2 report was not received by ERIC and is unavailable.  
PUB TYPE Reports - Descriptive (141)  
EDRS PRICE MF01/PC02 Plus Postage.  
DESCRIPTORS Accountability; College Students; \*Disclosure; Eligibility; Error Patterns; Federal Programs; \*Financial Aid Applicants; Grants; Higher Education; Management Information Systems; \*Models; Program Administration; \*Quality Control; Research Methodology; Statistical Analysis; \*Student Financial Aid  
IDENTIFIERS \*Basic Educational Opportunity Grants; \*Fraud

## ABSTRACT

Parameters and procedures for developing an error-prone model (EPM) to predict financial aid applicants who are likely to misreport on Basic Educational Opportunity Grant (BEOG) applications are introduced. Specifications to adapt these general parameters to secondary data analysis of the Validation, Edits, and Applications Processing Systems (Task 1) are also provided. After briefly discussing three approaches to error-prone modeling that are used at the state and federal levels, four types of validation applicants are identified: exact reporters, overclaimers, underclaimers, and validation dropouts. The research method, automatic interaction detection (AID), is described, and two ways to report the results of an AID analysis (a tree diagram and tabular format) are illustrated. Additional methodological issues are covered: the three study samples (working sample, replication sample, and nonvalidation sample); selecting and coding predictor variables; and the effectiveness of the model as indicated by a contingency coefficient. Flow charts illustrate the steps of the study, as well as the classification of applicants into validation types. Also outlined is a strategy for analyzing the quality control sample, which is derived from the 1978-1979 applicant file. Appendices detail the statistical analysis method. (SW)

\*\*\*\*\*  
\* Reproductions supplied by EDRS are the best that can be made \*  
\* from the original document. \*  
\*\*\*\*\*

HE

G-170

QUALITY CONTROL ANALYSIS OF SELECTED ASPECTS  
OF PROGRAMS ADMINISTERED BY THE BUREAU OF  
STUDENT FINANCIAL ASSISTANCE

Task 1 and Quality Control Sample  
Error-Prone Modeling Analysis Plan

March 31, 1980

Prepared for:

Bureau of Student Financial Assistance  
U.S. Office of Education  
Department of Health, Education and Welfare

Contract No. 300-79-0742

U.S. DEPARTMENT OF EDUCATION  
NATIONAL INSTITUTE OF EDUCATION  
EDUCATIONAL RESOURCES INFORMATION  
CENTER (ERIC)

- ☒ This document has been reproduced as  
received from the person or organization  
originating it.  
☐ Minor changes have been made to improve  
reproduction quality.

- Points of view or opinions stated in this docu-  
ment do not necessarily represent official NIE  
position or policy.



APPLIED  
MANAGEMENT  
SCIENCES

ED264754

AE 018 927

G-170

TASK 1 AND QUALITY CONTROL SAMPLE  
ERROR-PRONE MODELING ANALYSIS PLAN

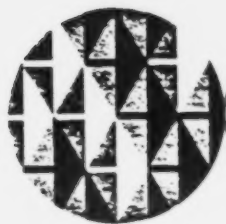
March 31, 1980

Prepared for:

Bureau of Student Financial Assistance  
U.S. Office of Education  
Department of Health, Education and Welfare

Contract No. 300-79-0742





APPLIED  
MANAGEMENT  
SCIENCES

962 Wayne Avenue • Suite 701 • Silver Spring, Maryland 20910  
Telephone 301 585-8181

March 31, 1980

Ms. Sylvia Cowan  
Project Officer  
Division of Quality Assurance  
Bureau of Student Financial Assistance  
Room 4673, ROB #3  
7th & D Streets, S.W.  
Washington, D.C. 20202

Reference: "Task 1 and Quality Control Sample Error-Prone Modeling  
Analysis Plan", Contract No. 300-79-0742 (AMS No. G-170)

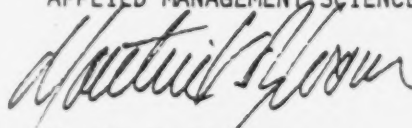
Dear Ms. Cowan:

We are pleased to submit eight (8) copies of the above-referenced Analysis Plan, which contains analysis specifications for conducting the error-prone modeling component of the Task 1 analysis. This Analysis Plan was prepared by Dr. Pedro Saavedra with the assistance of Ms. JoAnn Kuchak and Dr. Maureen Murphy.

We would like to schedule a meeting with you to discuss this plan prior to implementation. If you have any questions related to this plan please do not hesitate to contact me or Dr. Saavedra.

Sincerely,

APPLIED MANAGEMENT SCIENCES, INC.

  
Martin V. Bloom  
Executive Vice-President

MVB/rcc  
Enclosures

# TABLE OF CONTENTS

| <u>Chapter</u>                                                               | <u>Page</u> |
|------------------------------------------------------------------------------|-------------|
| 1. Introduction . . . . .                                                    | 1           |
| 2. Overview of Error-prone Modeling . . . . .                                | 2           |
| 3. Research Objectives and Study Groups . . . . .                            | 5           |
| 4. Research Issues . . . . .                                                 | 6           |
| 5. Overview of the Method . . . . .                                          | 7           |
| 6. Methodological Approach for the Task 1 Error-prone Model . . . .          | 13          |
| 6.1 Sample . . . . .                                                         | 13          |
| 6.2 Dependent variable . . . . .                                             | 15          |
| 6.3 Selection of the predictor variables . . . . .                           | 17          |
| 6.4 Coding the predictor variables and<br>handling of blank fields . . . . . | 21          |
| 6.5 Sequential search for applicant groups. . . . .                          | 22          |
| 6.6 Replication analysis . . . . .                                           | 22          |
| 6.7 Interpretation of resulting groups . . . . .                             | 22          |
| 6.8 Assessing the effects of validation for each group . . . .               | 23          |
| 6.9 Effectiveness of the model and comparison with PEC . . . .               | 23          |
| 7. Analysis of the Quality Control Sample . . . . .                          | 26          |
| 8. Expected results . . . . .                                                | 27          |
| Technical Appendix . . . . .                                                 | A1          |

# LIST OF EXHIBITS

| <u>Exhibit</u> | <u>Title</u>                                                                                                           | <u>Page</u> |
|----------------|------------------------------------------------------------------------------------------------------------------------|-------------|
| 1.             | AID Tree Diagram With Quantitative Dependent Variable . . . . .                                                        | 10          |
| 2.             | AID Tree Diagram With Dichotomous Dependent Variable: Supplemental Security Income Error-Prone Profile Study . . . . . | 11          |
| 3.             | SSI Group Summary Chart . . . . .                                                                                      | 12          |
| 4.             | BEOG Study Potential Group Summary Chart . . . . .                                                                     | 14          |
| 5.             | Flowchart Depicting Classification of Applicants into Types . . . . .                                                  | 18          |
| 6.             | Procedural Flowchart . . . . .                                                                                         | 24          |

## TASK 1 AND QUALITY CONTROL SAMPLE ERROR-PRONE MODELING ANALYSIS PLAN

### 1. INTRODUCTION

The Bureau of Student Financial Assistance has contracted with Applied Management Sciences, Inc. to provide analyses of pre-existing information on Basic Educational Opportunity Grant (BEOG) applicants that will yield findings for management of the program as well as findings specific to the characteristics of students who misreport on their applications. Three separate analytic substudies plus several optional tasks address BSFA's information needs. They include:

- o Secondary Data Analysis of Validation, Edits and Application Processing Systems (VEAPS)—Task 1,
- o IRS-BEOG Data Tape Match (IRS)—Task 2, and
- o End-of-Year Report on 1978-1979 BEOG Recipients (EYR)—Task 3.
- o Various analyses pertaining to the Quality Control Sample—Optional Tasks.

In addition, a component requested by the BSFA cuts across Tasks 1 and 2, and the optional tasks. This involves the development of an error-prone model (EPM) for the identification of applicants who are likely to be misreporting on their BEOG applications.

The purpose of this analysis plan is two-fold:

- o To present general parameters and procedures for developing an error-prone model. This general approach will form the basis for specific adaptations.
- o To present specifications for adapting these general parameters to Task 1 and the Quality Control Sample. At the appropriate time, parameters will be specified for Task 2.

This analysis plan will provide an overview of error-prone modeling, discuss research objectives and study groups, formulate research issues, provide an overview of the sequential analysis method to be used in the study and describe in detail the steps to be followed in carrying out the development of the model for Task 1 and the Quality Control sample.

Sections 2 through 6 will refer to the Task 1 derived error-prone model, though some of the objectives, issues and methods discussed

therein will also apply to the Quality Control study. Section 7 will briefly discuss the Quality Control sample analysis, while Section 8 will discuss expected results.

The expected results of this study will be:

- o A model useful for future adaptations to various needs of the BEOG program.
- o Information for management improvements specifically related to
  - selection of applicants for validation,
  - the edit system,
  - reduction of drop-outs among students entitled to funds, and
  - a global quality assurance strategy.

## 2. OVERVIEW OF ERROR-PRONE MODELING

In April, 1979, a National Conference on the Utilization of Characteristics Profiles as a Workload Planning Technique was held in Arlington, Virginia. This conference was devoted primarily to the development and utilization of error-prone models for various types of aid programs. The mere fact that this conference took place indicates that the development of adequate error-prone models has become an important concern for various agencies. The need to determine which cases are most likely to be misreporting so that validation procedures can be instituted for those cases is likely to be present for any program engaged in the disbursement of public funds on the basis of stated need and qualifying conditions.

If one is to investigate the characteristics of error-prone applicants, one must first define what constitutes an error. This generally refers to a discrepancy between what the applicant has reported and some independently ascertained determination of the "true facts" the applicant should have reported. This can be done through a validation procedure or through cross-checking with some other source of information about the applicant.

The Task 1 analysis will assume that validation has succeeded in establishing what the applicant should have reported, and that any

discrepancy between selection and payment transactions is to be treated as an error. There are certain obvious dangers to this assumption. There is no way to determine how conscientious financial administrators have been in validating applicants. Various kinds of misreporters (such as those who have unreported income sources) might find it easier to withstand validation efforts than other kinds of misreporters. Legitimate changes in circumstances would be classified as misreports if a supplemental form were not filed.

All of the drawbacks listed above are inevitable limitations of the study. On the other hand, some can in a way aid the purpose of the error-prone model. Since one purpose of the study is to recommend which candidates should be validated, the restriction of the definition of error to errors which can be discovered through validation, would be an asset rather than a hindrance. On the other hand, failure to validate on the part of financial aid administrators might obscure the results by making some misreporters appear to be accurate reporters. The Quality Control and the Task 2 analyses will, of course, lack some of these limitations while presenting some of its own.

Three major approaches to error-prone modeling have been used by various state and federal agencies. Each has its advantages and disadvantages and each is best suited to different kinds of situations.

The first method, used by the Welfare Departments of South Carolina and the District of Columbia, uses discriminant analysis to obtain a formula which assigns to each case a score. The higher the score, the more likely it would be that the applicant is misreporting. Thus if the agency wanted to select applicants most likely to misreport, it would simply select those applicants to whom the formula assigned the highest scores. If the agency wanted to insure selection of a group with a given percentage of misreporters in it, the selection of all applicants above an easily determined score would accomplish this.

The major drawback of this method is that it ordinarily assumes that a variable will affect all applicants in the same way. If it turns out

that estimating taxes makes a big difference for dependent applicants, but not for independent applicants, discriminant analysis will fail to take this into account. Thus it could easily fail to detect some important combinations of variables which could predict error-proneness. In addition, discriminant analysis would not point to specific areas where an applicant is likely to be misreporting. Of course, separate analyses could be conducted to predict misreporting for each specific field, but this method would lack parsimony and would be difficult to interpret.

The second approach, used by the State of New Hampshire to identify error-prone cases in its MEDICAID program, attempts to define a single group most likely to exhibit a high degree of misreporting. Depending on the size of the group, every member or a certain percentage of this group (and this group only) would be validated. Where only a small proportion of all the cases can be validated, and the principal objective is to maximize the savings in actual disbursements from the cases actually validated, this method can be very effective. On the other hand, this approach is likely to overlook groups whose error rate might approach that of the selected group, and which might require a different type of corrective measure. The BEOG program, with the use of edits in addition to validation, a high drop-out rate among applicants selected for validation and the dual concern for deterrence as well as savings in disbursements to validated applicants, requires a different approach.

The third approach has been used by the State of West Virginia AFDC program and by the Supplemental Security Income program of the SSA. It essentially divides the applicant population into mutually exclusive groups which differ as much as possible from each other in either the mean discrepancy in expected disbursement or in the rate of misreporting (or of various types of misreporting). This method has the advantage of being able to describe each group separately in terms of type of error, and thus to prescribe different types of corrective action for each. It has the further advantages of taking into account effects which apply to only part of the population, and of producing results which can be

expressed in simple terms. This method, sometimes called classification analysis, sequential structure search or automatic interaction detection (AID) is the one which will be used in this study.

### 3. RESEARCH OBJECTIVES AND STUDY GROUPS

The Pre-Award Validation Study and other aspects of the present Task 1 analysis have examined in global terms the misreporting and corrections behavior of the applicant population. What this analysis plan proposes is the development of error-prone profiles which will allow the BSFA to infer the individual applicant's misreporting behavior based on his profile at the time he was selected for validation.

At this point it is appropriate to introduce the terms group and type as they will be used in the study. The term group will refer to a classification of applicants based on data available at the time of selection for validation. This classification will be the major result of the study, and will be so defined that every regular applicant selected for validation will belong to one and only one group. Since the definitions of the groups will use no post-validation information, any applicant who could have been selected for the random validation procedure (i.e. any regular applicant with an EI under 1600) could also be classified into one of the resulting groups at any given transaction.

The term type will refer to the misreporting and post-validation behavior of each applicant. The conceptual definition of the four types into which validation applicants will be classified follows (operational definitions will be provided in section 6.2)

- o Type I: Exact Reporters. Validation confirms the information they provided.
- o Type II: Over-claimers. Misreported to their advantage.
- o Type III: Under-claimers. Misreported to their disadvantage.
- o Type IV: Drop-outs. Did not go through with validation.

Thus the major procedural objective of the study can be restated as follows:

To separate applicants into groups based on their characteristics prior to validation in such a way as to make the groups differ from each other as much as possible in terms of percentage of members of each group belonging to each type.

Each of the resulting groups can then be examined separately in order to determine its pattern of misreporting and post-validation behavior. The following substantive objectives will be pursued:

- To infer the pattern of misreporting and post-validation behavior for each applicant based on information available at selection.
- o To examine the characteristics of groups with a high proportion of applicants who failed to re-enter the system after selection.
- o To examine the effects of validation on members of each group separately.
- o To suggest further analysis and/or corrective action for each group separately.
- o To examine the effectiveness of the existing PEC, particularly in relation to the resulting error-prone model.

All of the above objectives lead to the main purpose of the study:

- o To suggest a strategy for validation that will most effectively minimize inappropriate disbursements while insuring applicants of the financial assistance to which they are entitled.

#### 4. RESEARCH ISSUES

The following research issues will be addressed by the error-prone modeling study.

1. How can applicants best be classified into groups using only information available at or prior to selection for validation, so that the groups differ as much as possible in terms of their post-selection characteristics.
2. What kinds of errors and in what fields are most prevalent among members of each group?
3. What were the results of validation for applicants who did not drop out, but belonged to groups with a high drop-out rate?

4. What is the mean discrepancy in expected disbursement between the selection transaction and the payment transaction for those members of each group who did not drop out?

5. How effective are the PEC when compared with the model emerging from the classification analysis?

6. What sort of further analysis and/or corrective actions seem to be appropriate for each of the resulting groups?

7. What would have been the subsequent corrections and reporting behavior for members of each group had they not been selected for validation?

#### 5. OVERVIEW OF THE METHOD

Automatic Interaction Detection (AID) is a research technique developed at the Institute for Social Research at the University of Michigan. ISR has developed two programs which will conduct this type of analysis: AID3 which accepts continuous dependent variables, and THAID which accepts categorical dependent variables. This software was used by the Supplemental Security Income program in their development of error-prone profiles. In addition, the West Virginia AFCD program used a software package closely related to THAID. In the subsequent discussion the term AID will refer to the general technique rather than to a specific program or package.

Techniques such as discriminant analysis or multiple linear regression make the assumption that a given effect will apply equally to all members of the population. The result of either of these techniques is an equation which is meant to predict the dependent variable for every member of the population. AID, on the other hand, does not make the assumption that a predictor variable will affect the dependent variable in the same way for all cases. Instead, AID starts by breaking up the sample into two subgroups selecting that split which can be defined in terms of one predictor variable which will produce groups which differ from each other on the dependent variable as much as possible. Each subgroup is then split separately, allowing for different predictor variables to split different previously formed subgroups.

This allows, for example, for the possibility that non-taxable income might be a good indicator of error-proneness among single independent applicants, but not among married independent applicants. Another possibility that linear methods might miss could be that among high income applicants, low taxes paid might be an indicator of error-proneness, but among low income applicants high taxes paid might be related to misreporting. This type of situation where a variable predicts differently for one group than for another is known as an interaction effect. Ordinarily, discriminant analysis and multiple linear regression do not take interaction effects into account. AID specifically tries to identify groups (using various combinations of variables) which will differ as much as possible on some criterion variable.

AID takes one dependent variable, which may be categorical (such as type of applicant) or quantitative (such as discrepancy in expected disbursement, expressed in dollars). Limitations on the number of predictor variables depend on the package, the number of categories per variable and a variety of other factors (THAID limits predictors to 40 per run). In practice, however, several runs can be combined to circumvent the limitations, in some cases running an entire analysis on a subgroup formed by a previous run.

Predictor variables can be monotonic (where the sample or any subgroup can only be subdivided into high and low groups based on some cut-off point) or free (used for categorical, as opposed to quantitative, predictors where any combination of values can be used to split the groups). In either case predictor variables must be coded in terms of a small number of possible values (no greater than 10 for THAID).

Out of the many possible splits defined by predictors, AID selects the one which will divide the sample into two groups as different from each other as possible (the operational definition of "as different from each other as possible" is given in the technical appendix). The process then continues for each of the two groups into which the original sample was split. When a group becomes very homogeneous, cannot be further

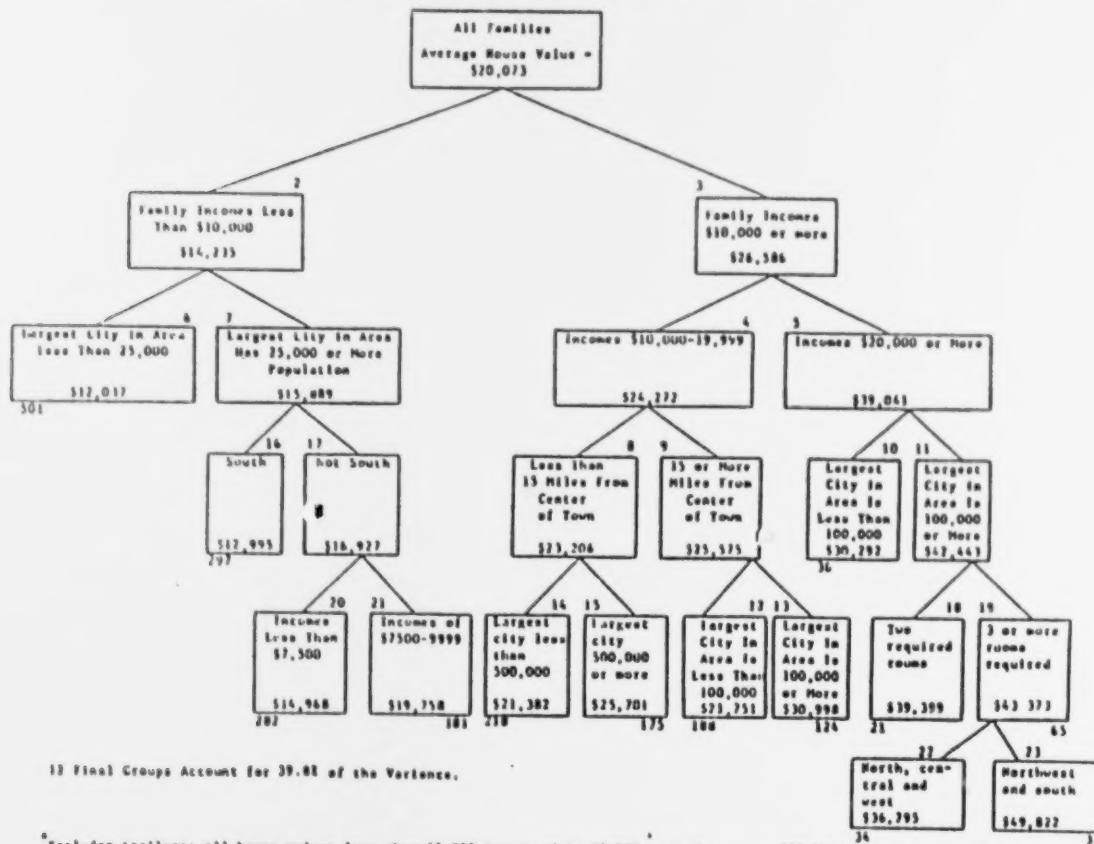
split using the predictor variables available, or would yield subgroups under a certain size if it were split, it then becomes one of the applicant groups. If an applicant group is both large and heterogeneous, it would be an indication that some additional predictors should be sought and included in the analysis.

Because AID investigates many possible combinations of variables, at times it produces results which are specific to a given sample. Two questions may be asked pertaining to the groups which emerge from an AID analysis: a) Do the groups have the same characteristics in the population as they appear to have in the sample? and b) Do the groups constitute an optimal classification of the population if one is trying to predict the dependent variable? In order to answer either question one needs to use a second sample randomly drawn from the same population. The first question can be answered by checking whether the subgroups produced by AID from the first sample have similar characteristics in the second sample. The second question requires that a separate analysis be conducted for the second sample. Unfortunately, it is quite possible that one would obtain a different solution if variables which are highly interrelated are used (this is similar to the problem of multicollinearity in multiple linear regression). The question of whether a solution is the best possible, however, is of secondary importance to whether the classification which emerges is effective in predicting error-proneness and in both saving funds and insuring fairness.

It is customary to report the results of an AID analysis either through a tree diagram or in tabular format. Exhibit 1 presents a tree diagram taken from Searching for Structure, a book by Sonquist, Baker and Morgan which describes the AID procedure for quantitative dependent variables. Exhibit 2 presents a tree diagram derived from the SSI study of error-prone profiles. Note that the SSI study uses the term profile in place of group. The SSI study used small regional samples (the example presented is one of many analyses conducted) so the number of applicant groups that resulted from the analysis was small. The same results could have been presented in tabular form (Exhibit 3).

EXHIBIT 1: AID TREE DIAGRAM WITH QUANTITATIVE DEPENDENT VARIABLE\*

Figure A5-1  
House Value (1970) by Family and Location Factors<sup>a</sup>



<sup>a</sup>Excludes trailers; all house values less than \$5,000 increased to \$5,000; and those over \$75,000 reduced to \$75,000.

\*Taken from Searching for Structure by J.A. Sonquist, E.L. Baker and J.N. Morgan, Institute for Social Research: Ann Arbor, Mi., 1974.

EXHIBIT 2: AID TREE DIAGRAM WITH DICHOTOMOUS DEPENDENT VARIABLE; SUPPLEMENTAL SECURITY INCOME PROGRAM  
ERROR-PRONE PROFILE STUDY

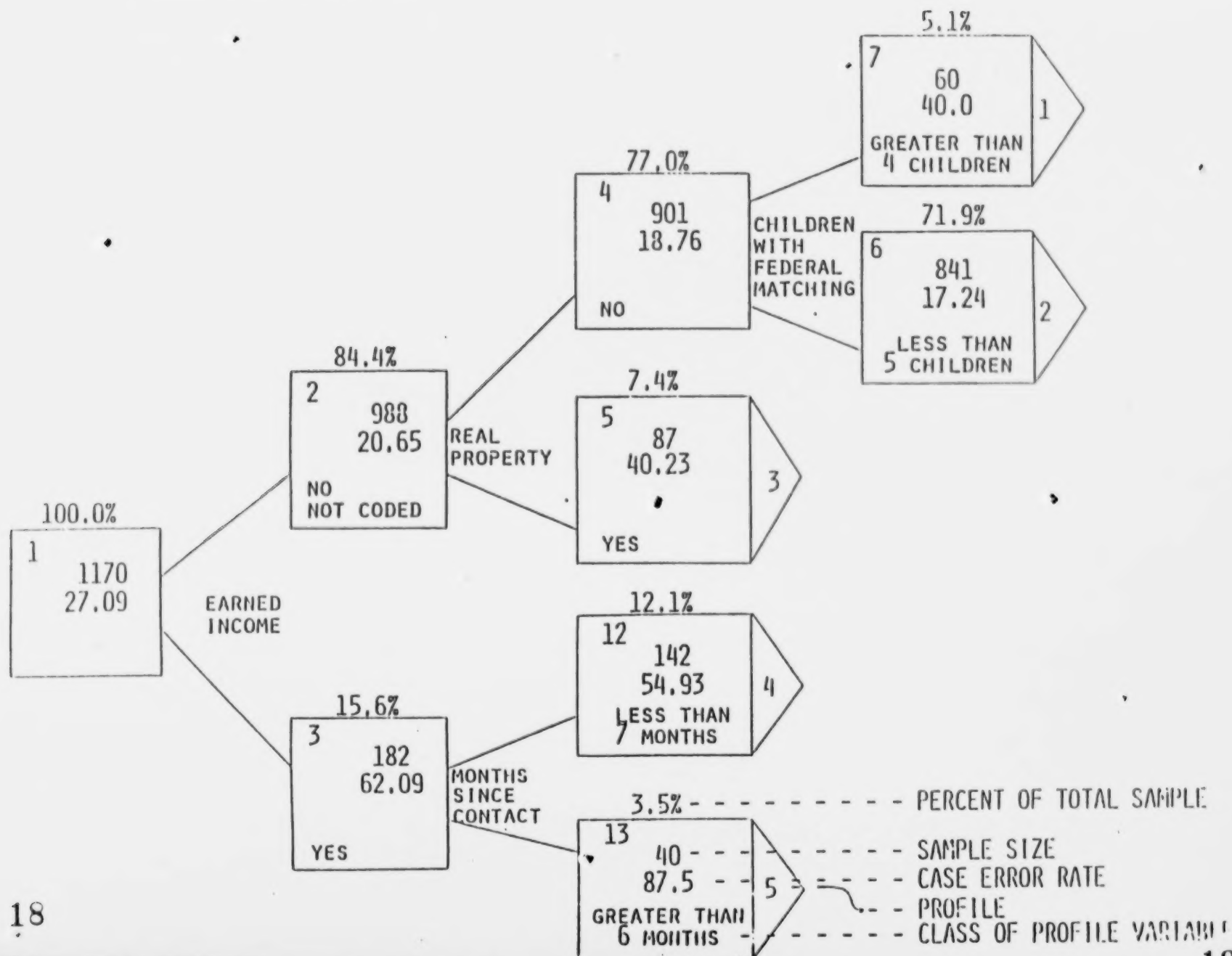


EXHIBIT 3: SSI GROUP SUMMARY CHART

| STEP 1                                           | STEP 2                                            | STEP 3                     | GROUP | PERCENT<br>OF CASES | CASE ERROR<br>RATE |
|--------------------------------------------------|---------------------------------------------------|----------------------------|-------|---------------------|--------------------|
| EARNED<br>INCOME<br><br>NO<br>OR<br>NOT<br>CODED | REAL<br>PROPERTY<br><br>NO                        | 5 OR MORE<br>CHILDREN      | 1     | 5.1                 | 40.00              |
|                                                  |                                                   | LESS THAN<br>5<br>CHILDREN | 2     | 71.9                | 17.24              |
|                                                  | REAL<br>PROPERTY<br><br>YES                       |                            | 3     | 7.4                 | 40.23              |
| EARNED<br>INCOME<br><br>YES                      | LESS THAN<br>7<br>MONTHS<br>SINCE LAST<br>CONTACT |                            | 4     | 12.1                | 54.93              |
|                                                  | 7 OR<br>MORE<br>MONTHS<br>SINCE<br>LAST CONTACT   |                            | 5     | 3.5                 | 87.50              |

In the BEOG analysis for Task 1 the dependent variable will consist of the four types. Exhibit 4 indicates what the results of this analysis might look like. The three samples will be discussed in section 6.1, but here the important thing is to note that each dotted line in each of the first four columns represents a split of either the sample (Col. 1) or a subsample resulting from a prior split (Cols. 2, 3 and 4). Each split would use only one predictor variable, but each of the final groups would require up to four predictor variables to define.

More detailed specifications of the analysis as it pertains to Task 1 follows in section 6.

#### 6. METHODOLOGICAL APPROACH FOR THE TASK 1 ERROR-PRONE MODEL

6.1 Sample. Three different samples will be used in this study. The first two samples will be drawn from the population of applicants randomly selected for validation. The third sample will be drawn from the population of applicants not selected for validation.

The first sample will be referred to as the working sample. This sample will be drawn from among those applicants randomly selected for validation by a given date. The reason for placing a cut-off date by which an applicant must have been selected for validation for inclusion in the sample is the need to distinguish between cases that are pending and cases that have failed to re-enter the system. The date will be arrived at based on an examination of both the 1978-1979 and the 1979-1980 files. The date will be chosen so as to minimize the proportion of applicants with no post-selection transaction who have not made it to the recipient file by the time the sample is drawn but make it after the drawing of the sample. Such applicants would be incorrectly classified as drop-outs. It is expected that such a date would be three to six months prior to the drawing of the sample. Sample size will be five to ten thousand, depending on the size of the population randomly selected for validation by the chosen date. Sample size must not exceed fifty percent of the population.

# EXHIBIT 4: BEOG STUDY POTENTIAL GROUP SUMMARY CHART

| Classification Results By Type And Sample |        |        |        |           |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|-------------------------------------------|--------|--------|--------|-----------|----------------|------|---|----|-----|--------------------|---|------|---|----|-----------------------|----|---|------|---|
| Step 1                                    | Step 2 | Step 3 | Step 4 | Group No. | Working Sample |      |   |    |     | Replication Sample |   |      |   |    | Non-Validation Sample |    |   |      |   |
|                                           |        |        |        |           | N              | Sam. | I | II | III | IV                 | N | Sam. | I | II | III                   | IV | N | Sam. | I |
|                                           |        |        |        | 1         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 2         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 3         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 4         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 5         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 6         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 7         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 8         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 9         |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 10        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 11        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 12        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 13        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 14        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 15        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |
|                                           |        |        |        | 16        |                |      |   |    |     |                    |   |      |   |    |                       |    |   |      |   |

Each step represents a split of the sample or of previously split subsamples.

- I. Exact Reporters
- II. Over-claimers
- III. Under-claimers
- IV. Drop-outs

The second sample will be referred to as the replication sample. It will be drawn in exactly the same way as the working sample, but it will not overlap in its membership with the working sample. In the extreme case of a small population randomly selected for validation by the chosen date, the working and replication samples will each contain half the population. The use of a replication sample minimizes the possibility of conclusions based on unreliable results.

The third sample will be referred to as the non-validation sample and will be drawn from the population of applicants who were not selected for validation. This sample will be drawn by selecting every fiftieth record provided the record met the criteria for random selection by the date chosen for the first two samples until the sample size equals that of the working sample. Thus the sampling procedure will not only choose applicants, but will select a transaction for each applicant. For analysis purposes this transaction will be treated in the same way the selection transaction is treated for the other two samples. The non-validation sample will be used to address research issue #7.

6.2 Dependent variable. As mentioned earlier AID can use either a categorical or a quantitative dependent variable. The best candidate for a quantitative dependent variable seems to be discrepancy in expected disbursement between the selection transaction and the latest payment transaction. Use of this discrepancy as a dependent variable in the AID analysis would have the advantage of differentiating groups with many applicants having small discrepancies from those where the discrepancies are major. On the other hand, the use of a quantitative variable would mean either dropping applicants who failed to re-enter the system after selection from the analysis or treating them as if their post-validation expected disbursement was zero. The first alternative would ignore a large proportion of the population, while the second would probably lead to misleading results.

The alternative is to create a categorical variable based on the four types mentioned in section 3. This still leaves open the question of what to do with applicants who did not have post-selection transactions.

The assumption will be made that if an applicant was selected for validation, had no post-selection transaction and then became a recipient, the information he provided at selection was essentially accurate. This, of course, assumes the FAA carried out the validation procedure—an unavoidable limitation of the study.

The drop-out type presents problems of its own. An applicant who filed no post-selection transaction and who did not become a recipient on the basis of the selection transaction will be classified as a drop-out. Applicants whose latest transaction was a rejection will also be considered drop-outs for analysis purposes. This allows for the possibility that an applicant may be in the process of filing a new transaction and will be classified inappropriately as a drop-out. On the other hand, given the requirement that all applicants in the sample should have been eligible several months prior to the study, such applicants should constitute a very small percentage of each sample.

A third problem is presented by applicants who become ineligible after selection. If an applicant has not been paid on selection or on a post-selection transaction and appears as ineligible on his latest transaction, that latest transaction will be treated as a payment transaction with an expected disbursement of zero.

Note that an applicant could have been paid based on a transaction prior to the selection transaction. Since his classification into one of the four types is based on his selection transaction, he could still be classified as belonging to any of the four types (i.e., payment prior to selection is ignored).

Operational definitions for each of the four types follow:

Type I: Exact reporters. Applicants who meet one of two conditions:

- a) Were paid on the selection transaction.
- b) Were paid on a post-selection transaction, and their latest expected disbursement is within fifty dollars of their calculated expected disbursement at selection.

Type II: Over claimers. Applicants who misreported to their advantage (i.e. were payed on a post-selection transaction and whose latest expected disbursement is at least fifty dollars lower than their calculated expected disbursement at selection).

Type III: Under-claimers. Applicants who misreported to their disadvantage (i.e. who were payed on a post-selection transaction and whose latest expected disbursement is at least fifty dollars higher than the calculated expected disbursement at selection).

Type IV: Drop-outs: Applicants who meet one of two conditions:

a) Had no post-selection transaction and were not recipients based on the selection transaction.

b) Had post-selection transactions, were not paid on selection or on any post-selection transaction and had no transactions filed after a given date (to be determined) at least two months prior to the latest update of the recipient file.

All other cases will be considered pending, and will be excluded from the analysis but their characteristics will be reported separately.

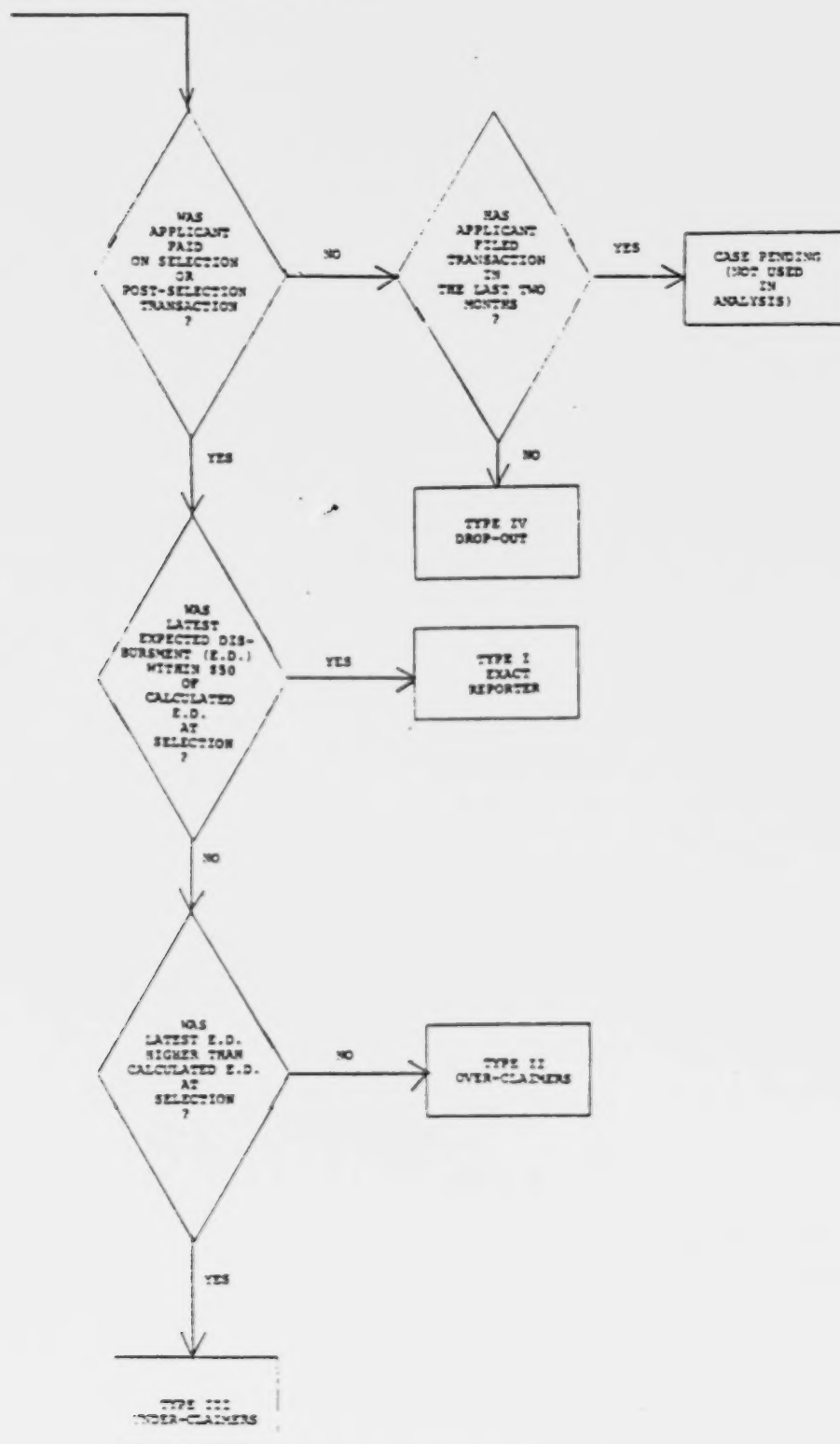
Exhibit 5 presents this information in a different form.

It will be noted that this classification does not distinguish between fields which might have produced discrepancies in expected disbursement. It has been the experience of the SSI study that different groups with high error rates will differ in terms of which fields produced the discrepancies in expected disbursement.

6.3 Selection of the predictor variables. This is possibly the most critical aspect of the analysis. A preliminary list of variables will be drawn from those used in previous studies, from recommendations made by persons familiar with the BEOG program and from variables which correspond to some that have been found useful in the development of error-prone profiles by other agencies.

A preliminary coding of each variable being considered as a predictor variable will be conducted and contingency tables reporting the number of cases acquiring each value and the percentage of such cases which fit each type will be produced. The decision to include a variable as a predictor variable in the analysis can be based on several factors:

EXHIBIT 5: CLASSIFICATION OF APPLICANTS INTO TYPES



- o Variables which appear to predict by themselves should almost always be included.
- o Variables representing critical fields should be included.
- o Variables related to PEC should be included, but this may be done after other variables have been tried first.
- o Variables which are highly interrelated should not be used at the same time. AGI and total income (AGI plus NTI) would be highly interrelated, so one or the other should be used, but not both. If one does not yield productive results, the other can be substituted in a subsequent analysis.

It is possible that a group will be found which cannot be split by any of the predictor variables available, but which is both large and heterogeneous. The presence of such a group would indicate the desirability of incorporating new predictor variables into the analysis, at least for that predictor alone.

Predictors can be classified into ahistorical, historical and mixed. A historical predictors refer to the selection transaction without regard for any other transactions filed by the applicant (e.g. AGI, dependency status or MDE source). Historical predictors refer to what happened prior to the selection transaction (e.g. whether AGI was ever corrected, number of prior transactions, whether applicant was ever ineligible or discrepancy between highest and lowest EI prior to validation). Applicants selected on their first transaction should receive the same values on historical variables. Mixed variables could refer to the present or prior transactions (e.g. was household size every left blank or highest EI obtained in any transaction). The only historical variable used in the first preliminary analysis conducted will be number of prior transactions. Once this variable is used to define a group, historical variables will be incorporated into the analysis for groups including applicants with multiple pre-selection transactions.

The following is a preliminary list of possible predictor variables:

- o AGI
- o NTI
- o Household Size

- o Dependency Status
- o Year in School
- o Marital status
- o Persons in post-high school education
- o Social Security Benefits present?
- o Date of first application
- o Geographical region
- o Number of dependency criteria met
- o Granted permission for information to be passed to state agency
- o Veterans Benefits present?
- o Source of tax figures
- o MDE source
- o Ratio of father or applicant's income earned to net assets
- o Ratio of mother or spouse's income earned to net assets
- o Itemized deductions
- o Medical and dental expenses
- o Casualty or theft losses
- o Home ownership present?
- o Investments
- o Cash and bank accounts
- o Business assets reported?
- o Farm assets reported?
- o Applicant's resources
- o Number of prior transactions
- o Met at least one PEC
- o Individual PEC variables
- o Corrected AGI
- o Corrected Household Size down in response to comments
- o Corrected Taxes Paid since previous transaction
- o Corrected other critical fields
- o Changed dependency status
- o Changed source of tax figures
- o Household Size ever blank
- o Ever failed to sign

- o Ever rejected
- o Difference in EI from lowest obtained to highest
- o Ever omitted field subsequently reported and different from zero?

These are obviously not meant to be exhaustive, but are mere examples.

6.4 Coding of predictors and handling of blank fields. AID accepts predictors of two types: monotonic and free. All variables must be coded into categories, but monotonic variables are restricted in that the only permissible splits are those that divide the sample or subsample into high and low groups. Free variables allow any combination of values to define the split. Quantitative variables such as AGI are usually coded as monotonic while categorical variables such as MDE source are usually coded as free. It is important to avoid free variables with many possible values because the number of splits the computer has to examine for a free variable increases exponentially with the number of values.

A preliminary coding will be conducted for each variable being considered for inclusion in the analysis. Contingency tables will be prepared for each variable presenting the number of cases for each value and the percentage of those cases fitting each of the four types previously discussed. Examination of these contingency tables will then lead to recoding appropriately into a smaller number of categories.

One important methodological issue is how to treat cases where a field was left blank by the applicant. In some cases using the assumed value is probably appropriate (for example, there is probably no difference between applicants who left business assets blank and those who took the trouble to write a zero). On the other hand, in some instances a blank field can be a predictor of misreporting (e.g. omission of Household Size could be significant). There are two possible approaches to this problem: the variable could be coded as free, coding blank as an additional category, or the variable could be coded as monotonic using the assumed value for a blank field and treating whether the field was blank or not as an additional variable. Examination of the preliminary coding should precede deciding for one approach or the other.

6.5 Sequential search for applicant groups. The actual AID analysis will be conducted using the THAID program. This program can conduct the search on the basis of either of two different statistics: theta or delta. The fact that the types differ greatly in terms of the proportion of the population they represent makes delta the appropriate statistic. Various aspects of the analysis have been discussed in the previous section, so here one must merely add that the process will be one of trial and error in terms of the predictor variables that will be used and splits that might be forced. The end result will be a classification of all applicants into mutually exclusive groups that will differ from each other as much as possible. While the program only guarantees that the groups will differ on the dependent variable it has been the experience of other users of AID to derive error-prone profiles that groups defined by different branches of the tree tend to produce different patterns of error. The groups defined by the analysis will be used as the error-prone model here proposed. This analysis will directly address Research Issue #1.

6.6 Replication analysis. A groups by types contingency table will be produced for the working sample and separately for the replication sample. The delta statistic will be used to identify any groups where the distribution of the types differs across samples. If such a group is identified the split which led to its formation will be eliminated. The AID analysis will be then recalculated omitting the variable which caused the split.

6.7 Interpretation of resulting groups. Contingency tables will be prepared from the working sample by groups crosstabulated with the following variables (plus any others the analysis might suggest)

- o Each of the critical fields as reported in the selection transaction.
- o Each of the variables incorporated at any stage of the analysis whether it was used in a split or not.
- o Each of the PEC.
- o Discrepancies in each critical field between selection and payment transactions and differences in expected disbursement resulting from each discrepancy calculated only for applicants who did not drop out.

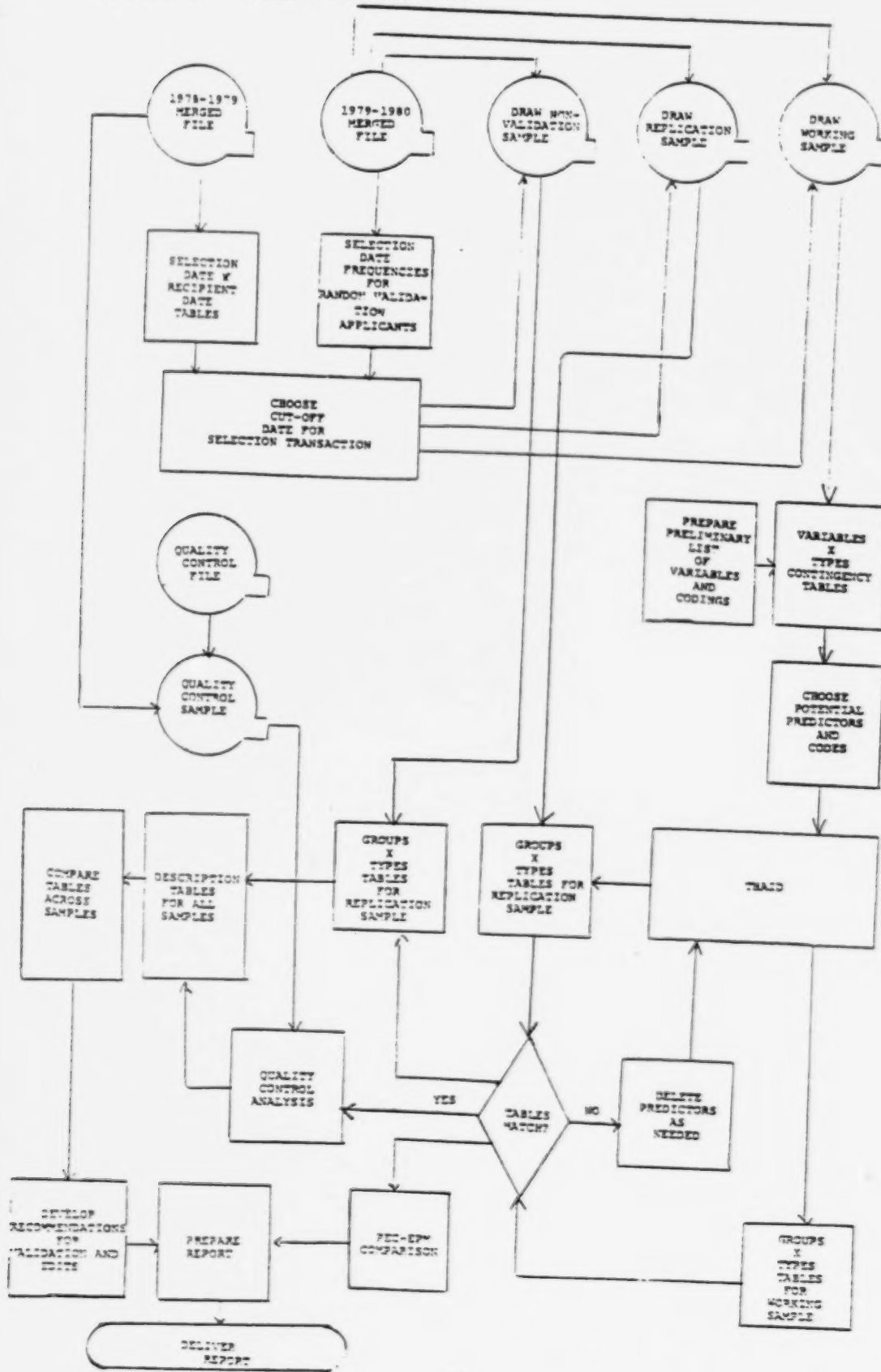
This aspect of the study will address Research Issues # 2, 3 and 4.

6.8 Assessing the effects of validation for each group. Types will be defined for the non-validation sample using the record chosen by the sampling procedure as if it were the selection transaction. A groups by types contingency table will be prepared and compared with those obtained from the working sample and the replication sample. Also discrepancies in expected disbursement in general and due to corrections to each critical field will be calculated and compared with those of the working sample. Research issues # 6 and 7 will be addressed by this aspect of the analysis. Exhibit 4 presents one of the tables which will be used for this part of the analysis.

6.9 Effectiveness of the model and comparison with PEC. One measure of the effectiveness of the model would be a statistic based on the groups by types contingency table (e.g. contingency coefficient or Cramer's V). Unfortunately such a measure could not be meaningfully applied to the PEC since an applicant can meet more than one PEC. However, if the drop-outs are eliminated from this analysis and discrepancy in expected disbursement from selection to the payment transaction is used as a dependent variable in a multiple linear regression with dummy predictors, the variance of the dependent variable can be partitioned into components predicted by each model, a component where the predictions overlap and a component unpredicted by either model. Such an analysis will be conducted.

The four resulting coefficients will add up to 1.0 and will provide an indication of the relative effectiveness of the error-prone model and the PEC. If the model fails to explain a large proportion of the variance it would not be worth pursuing. If the PEC explains a large proportion of the variance not explained by the model, then the model should be expanded to incorporate some of the PEC as predictors. If the model explains a large proportion of the variance and the PEC does not explain any additional variance, this would be an indication that the PEC should be dropped and validation procedures should be based on the error-prone model resulting from this study. Research Issue #5 will be examined by this aspect of the study.

# EXHIBIT 6: PROCEDURAL FLOW CHART



## 7. ANALYSIS OF THE QUALITY CONTROL SAMPLE

In order to incorporate corrections behavior into the Quality Control analysis, the applicants in the QC file will be picked out from the 1978-1979 file and the information from both files merged. The resulting applicant file will be used and referred to as the Quality Control Sample.

The Quality Control sample presents some unique advantages and some unique limitations when compared with the random validation sample. For this reason a separate analytic strategy must be developed for the Quality Control sample.

The most obvious difficulty presented by the QC sample is its size. While it is feasible to conduct an AID analysis with a sample of 2,309 cases, doing so in the present investigation would present the following difficulties:

- o The absence of a replication sample would present the risk of accepting results peculiar to the sample.
- o The number of groups that could be reliably identified with such a sample would be relatively small compared to the number that could be expected to emerge from the applicant population.
- o The population of BEOG applicants contains certain subgroups for which one would expect separate branches of the tree (independents, for example). The size limitations of any such subgroup might make further splits suspect.

In spite of these limitations an AID analysis will be attempted using the THAID program. The dependent variable will be defined using types just as with the random validation sample. The one change will be that in place of the drop-out type, applicants who misreported their dependency status will constitute a separate type. Results emerging from this analysis, however, must be viewed as tentative and will be derived primarily for comparison with the results obtained from the random validation sample.

A second approach will also be used to obtain maximum information from the QC sample. To the extent possible, the groups defined by the AID analysis on the random validation sample will be used to break down information obtained by the QC study. The reason why this process may

not be entirely successful is that the QC sample is based on 1978-1979 data while the random validation sample used to derive the groups is based on 1979-1980 data. Conceivably the discrepancies between the application forms or the arrangement of the files might make defining groups for the QC sample based on the random validation sample EPM inappropriate. On the other hand, the merging of two groups if needed or the formulation of assumptions concerning the match of 1978-1979 fields with 1979-1980 ones would allow the analysis to be conducted.

By matching the random validation EPM groups with applicant characteristics in the QC sample, the reason why certain groups exhibit a high rate of misreporting or a high drop-out rate might be discerned. In addition information can be obtained as to kinds of misreporting that validation failed to identify. In order to obtain this information, the following contingency tables will be prepared:

- o Groups by types (as defined for the QC sample)
- o Groups by discrepancies for each critical field
- o Groups by responses to questionnaire items

Further details of the QC sample error-prone analysis will of necessity depend on results of other analyses and will thus not be presented here.

## 8. EXPECTED RESULTS

As a result of the various analyses here proposed, this study will yield the following:

- o A classification system that will group applicants on the basis of a given transaction and can be used to predict the likelihood that validation will uncover misreporting on the part of the applicant or that the applicant will fail to re-enter the system if validation is required.
- o A determination for each group of the probable effects of validation and of the likely results had validation not been carried out.
- o A group by group discussion of the optimal strategy that seems to be suggested by the group profile in order to minimize misreporting or failure to re-enter by students who could become eligible were they to report correctly. In addition some group profiles might pose new questions which cannot be answered with the present data, but which deserve further investigation.

While the error-prone model does not directly examine the edit system, other than to incorporate corrections behavior into the set of predictor variables, the results will have implications for the edit system as much as for the validation system. If members of a particular group frequently misreport a given field, the incorporation of a new or more restrictive comment for members of the group might be just as effective as validation.

Possible recommendations or remarks that might be made for different groups follow. Please note that some of these might not apply to any of the resulting groups.

- o Validation is suggested. This group has a large proportion of over-claimers and their reports are not corrected without validation.
- o Members of this group appear to correct misreporting in response to comments. Validation is not necessary.
- o Misreporting for this group seems to be confined to \_\_\_\_\_. Developing new or more restrictive comments related to this field might solve the problem.
- o Under-claimers seem to be over-represented in this group. Recommending that applicants in this group see their FAA since they might be entitled to more funds might help correct this situation.
- o This group has a high drop-out rate after random selection for validation, but not where the applicant was not selected. There is no evidence of high rates of misreporting among group members that do not drop out and the Quality Control sample reveals no unusual pattern. Further investigation through interviewing applicants from this group and special attention to its results in the Task 2 investigation would be in order.
- o Both under-claimers and over-claimers predominate in this group. The Quality Control sample suggests members of this group either did not read the instructions completely or found them confusing. Edit messages suggesting a more careful reading of the instructions or contacting their FAA might help reduce misreporting.
- o Misreporters in this group were not identified through validation, but were identified in the Quality Control sample. Incorrect dependency status seems to be the most frequent cause of errors. A more thorough process than simple validation seems to be necessary for this group.

## TECHNICAL APPENDIX

I. The delta statistic. The program THAID, to be used in the main Task 1 analysis uses a nominal scale dependent variable. When THAID splits the sample or a subsample into two subgroups it calculate the delta statistic for each split.

$$\text{Delta} = \frac{N_1 \sum_{j=1}^G |P_j - p_{1j}| + N_2 \sum_{j=1}^G |P_j - p_{2j}|}{2(N - \frac{\sum N_j^2}{N})}$$

Where

$N_1$  = frequency in first subgroup

$N_2$  = frequency in second subgroup

$P_j$  = proportion of cses in unsplit group falling on the  $j^{\text{th}}$  code of the dependent variable

$p_{ij}$  = proportion of cases in  $i^{\text{th}}$  split subgroup falling on the  $j^{\text{th}}$  code of the dependent variable

$N$  = frequency in unsplit group

$N_j$  = frequency in unsplit group in  $j^{\text{th}}$  code

$G$  = number of possible values in the dependent variable.

THAID chooses that split which can be defined by a predictor variable which maximizes delta.

II. Partitioning the variance. For only those cases where a pre and post validation discrepancy in expected disbursement (Y) can be calculated two sets of dummy variables will be prepared. Set A will consist of  $n-1$  dummy variables where  $n$  is the number of terminal groups produced by the THAID analysis. For every group but the last  $A_j=1$  if a case belongs to the group, 0 otherwise. Set B will code dummy variables for each of the PEC.  $B_j=1$  if the case meets the corresponding PEC, zero otherwise.

Multiple linear regression will be used to partition the variance. The regression procedure will be used to calculate  $R_{Y A}$ ,  $R_{Y B}$  and  $R_{Y AB}$ .

- (1)  $1 - R_{Y AB}^2$  = unpredicted proportion of the variance
- (2)  $R_{Y AB}^2 - R_{Y A}^2$  = portion predicted by PEC, but not model
- (3)  $R_{Y AB}^2 - R_{Y B}^2$  = portion predicted by model, but not PEC
- (4)  $R_{Y A}^2 + R_{Y B}^2 - R_{Y AB}^2$  = portion of the variance accounted for by both models
  - (3) + (4) = proportion of variance explained by model.
  - (2) + (4) = proportion of variance explained by PEC.
  - (2) + (3) + (4) = proportion of variance that would be explained if both PEC and model were used.